Received: 21 November 2022 | Accepted: 1 November 2023

DOI: 10.1002/psp.2732

RESEARCH ARTICLE

WILEY

The migrant perspective: Measuring migrants' movements and interests using geolocated tweets

Johannes Mast[1] | Marta Sapena[1] | Martin Mühlbauer[1] | Carolin Biewer[2] | Hannes Taubenböck[1,3]

[1]German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Weßling, Germany

[2]Department of English and American Studies, Chair of English Linguistics, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

[3]Institute for Geography and Geology, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

Correspondence

Johannes Mast, German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Germany, Münchener Str. 20, Weßling 82234, Germany.
Email: johannes.mast@dlr.de

Funding information

European Union's Caroline Herschel Framework Partnership Agreement on Copernicus User Uptake (FPCUP), Grant/Award Number: FPA275/G/GRO/COPE/17/10042; Volkswagen Foundation, Programme "Momentum—Funding for Recently Tenured Professors", Grant/Award Number: 98662; German Federal Ministry of Education and Research, Grant/Award Number: 01LG2082C

Abstract

Geolocated social media data hold a hitherto untapped potential for exploring the relationship between user mobility and their interests at a large scale. Using geolocated Twitter data from Nigeria, we provide a feasibility study that demonstrates how the linkage of (1) a trajectory analysis of Twitter users' geolocation and (2) natural language processing of Twitter users' text content can reveal information about the interests of migrants. After identifying migrants via a trajectory analysis, we train a language model to automatically detect the topics of the migrants' tweets. Biases of manual labelling are circumvented by learning community-defined topics from a Nigerian web forum. Results suggest that differences in users' mobility correlate with varying interests in several topics, most notably religion. We find that Twitter data can be a flexible source for exploring the link between users' mobility and interests in large-scale analyses of urban populations. The joint use of spatial techniques and text analysis enables migration researchers to (a) study migrant perspectives in greater detail than is possible with census data and (b) at a larger scale than is feasible with interviews. Thereby, it provides a valuable complement to interviews, surveys and censuses, and holds a large potential for further research.

KEYWORDS

domain adaptation, human migration, mobility, NLP, social media, trajectories

1 | INTRODUCTION

Human migration is a complex phenomenon, and its study is still constrained by a lack of data (Kirchberger, 2021). Rather than being a precisely defined process, migration is part of the spectrum of human mobility patterns which range from occasional travel to permanent relocation (Willekens et al., 2016). In a globalized world, migration is intertwined with political, economic and cultural processes (McAuliffe & Ruhs, 2018), but how these processes factor into migrant decision-making is still not sufficiently understood. Important insights could be gained by better understanding the perspective of individuals (McAuliffe et al., 2018), those who decide to move as well as those who decide to stay (Schewel, 2020). Who are they, where are they, and what matters to them? A data source which can keep up with dynamic and international populations across borders, and inform about their views and experiences in changing environments, would be highly valuable (McAuliffe et al., 2018).

The same dynamism and variety of mobility patterns necessitate that the data not only provide high coverage, but can also be flexibly applied across space, time and topics. This is where traditional data sources, such as ethnographies, surveys and censuses, reach their limits (Rampazzo et al., 2021). Ethnographies and interviews provide rich detail, but rely on small samples (Rowe et al., 2021). Longitudinal surveys capture migration dynamics with detailed migration histories but at high data collection costs and efforts (Fussell et al., 2014). Demographic data from censuses or administrative sources can provide high geographic detail for large parts of the population, but are limited in detail, timeliness and coverage of migrants, and suffer from heterogeneity in their definitions of migration (Rampazzo et al., 2021; Spyratos et al., 2018; Willekens, 2019). Altogether, data suffers from cost and rigidity, and there appears to be a gap between detailed and large-scale information sources.

These information gaps could be reduced by types of data which have emerged from developments in information technologies (McAuliffe et al., 2018; Reips & Buffardi, 2012): Microblogs like Twitter provide geolocated social media (SM) data that is, to date, freely and globally available. The combination of text data and geolocation which these data offer enables the joint assessment of the interests and the mobility of users and reveals pathways towards a better understanding of migration.

2 | BACKGROUND: ANALYSIS OF MIGRANTS USING SM

In recent years, the number of people engaging with SM has grown substantially, and SM platforms are now recognized as spaces for socializing and reflecting on all aspects of everyday life (Townsend & Wallace, 2016; Zhu et al., 2022). Consequently, "social sensing" data have been proposed as a real-time and inexpensive way to measure social phenomena (Wang et al., 2019).

The opinions expressed on SM can inform about a wide variety of topics (Wang et al., 2019), making the usage of user-provided data a scientific trend in many fields of research (Kounadi & Resch, 2018). For migration studies, text-based social networks are an interesting data source, because they enable users to connect and exchange knowledge over long distances (Dekker & Engbersen, 2014), and have been considered central to international migration decision-making (Akanle et al., 2021). For qualitative studies, migration researchers have long recognized SM's potential to provide a unique insight into the interests and behaviours of migrants (Reips & Buffardi, 2012). But could migrant interests also be studied at large scales, using *big data* approaches?

The feasibility of large-scale studies of SM users has been demonstrated for a variety of text-based SM platforms, such as LinkedIn (Bastian et al., 2014), Facebook (Heidenreich et al., 2020) or Twitter (Giachanou & Crestani, 2016). Widespread methods for the analysis of such data include topic modelling (Calderón et al., 2020) and sentiment analysis (Giachanou & Crestani, 2016). When it comes to migration research, however, these studies have a blind spot. Khatua and Nejdl (2021) found that, so far, SM studies mostly explored public opinions *about* refugees and migrants (see, e.g., Heidenreich et al., 2020; Lee & Nerghes, 2018; Rowe et al., 2021), while the first person-perspectives *of* migrants have been neglected. This is a missing link of no small importance (McAuliffe et al., 2018): Subjective perceptions are key to migration decisions (Hoffmann et al., 2021), even if they do not always match the objective reality. In their aforementioned study, Khatua and Nejdl (2021) also demonstrate the feasibility of analyzing views and struggles of migrants on SM, but find that identifying migrants is challenging. Their approach of relying on migrants explicitly referring to themselves as such can detect only a miniscule, likely biased (Olteanu et al., 2019), subset of migrants, and is thus not able to make visible the needs and desires of the large number of migrants that are active on SM.

However, there are alternative methods which can be used to identify migrants. Language can be used as a proxy (Lamanna et al., 2018; Sîrbu et al., 2021), although limited by the dominance of English as a lingua franca (Kim et al., 2020). And where SM contains geoinformation, spatial analysis is a possibility. Various studies (Armstrong et al., 2021; Blumenstock, 2012; Fiorio et al., 2017; Gollin et al., 2021; Hawelka et al., 2014; Mazzoli et al., 2020; Spyratos et al., 2018; Zagheni et al., 2014) show that the identification of mobility patterns using geolocated SM data is feasible. Armstrong et al. (2021) find that the users identified by their method rarely represent migrants in the traditional sense but cover a wide range of mobile users including business travellers, tourists or global citizens. Therefore, SM data have the potential to be a common ground for the analysis and comparison of a range of mobile populations of various kinds which may not jointly appear in traditional data sets.

However, thus far, the analysis of movements from the point of view of geographical science and the analysis of texts from the point of view of social science have remained apart. Recently, works by Kim et al. (2020, 2021, 2022) have shown that migrants and natives can be distinguished by relating their geolocation to that of their friends (Kim et al., 2020), and characterized via the metadata and hashtags associated with their tweets (Kim et al., 2021). This demonstrates the potential of combining various facets of SM data, and while Hashtags are powerful labels for trending topics, an even greater wealth of information is contained in the main content of SM posts. Therein, user-created texts contain implicit and subjective information about a wide variety of topics and the users' attitudes towards them. Extracting this information from large data sets is only feasible with automated methods though, which is challenging because the features of informal Internet communication are different from traditional written text and often contain typographical errors (Nguyen et al., 2020).

Over the past years, substantial advances have been made in the field of natural language processing (NLP). Large general-purpose language models trained on very large corpora (Brown et al., 2020; Devlin et al., 2018) now excel at a variety of tasks, including multilabel text classification. This facilitates the automated extraction of complex topics even from large corpora in informal language (Kayastha et al., 2021). So far, the full potential of this has not been tapped in the context of migration, perhaps due to the lack of suitable training corpora or the aforementioned difficulty of identifying migrants. In this study, we demonstrate how these challenges can be overcome.

Our overarching goal is to develop a methodology which can harness geolocated SM data for migration research. Linking Nigerian Twitter users' mobility to their interest in certain topics, we provide a test case which explores the potentials and limitations of our method. As an alternative to the recent approach by Kim et al. (2021), we aim to provide a method that, given the availability of sufficiently rich data, can be flexibly adapted to various mobility forms and platforms, and captures general topics rather than specific hashtags, to match the variety of mobility forms and online conversations.

Our approach is twofold:

First, we use the geolocation information to identify mobile users via their trajectories. Second, we analyze the text content produced by these users to identify their topic preferences and compare them to those of stationary users.

The remainder of this paper is structured as follows. After an outline of the setting and the data source of our study (Section 3.1), we present the materials and methods used for the mobility analysis (Section 3.2.1) and the text analysis (Section 3.2.2). The results are likewise presented separately for the spatial analysis (Section 4.1) and the text analysis (Section 4.2). The discussion (Section 5) reflects on the key findings of the case study (Section 5.1) and discusses the implications for policy (Section 5.2), closing with a reflection on the limitations of the approach and ethical considerations (Section 5.3). Conclusions are drawn in Section 6.

3 | MATERIALS AND METHODS

3.1 | Study area and data source

Nigeria provides a relevant case study due to its cultural diversity, pivotal position in West Africa, and rapidly increasing young population (Central Intelligence Agency, 2022) with increasing penetration of mobile phones (Forenbacher et al., 2019). Today, as Kirwin and Anderson (2018) found, Nigerians comprise the vast majority of people in West Africa who are motivated to migrate. SM use is relatively low, but increasing. In January 2015, 13.6 million SM users were estimated in Nigeria, a 7% share of the country's total population (Kemp, 2015). In January 2019, that number had risen to 24 million, or 12% of the population (Kemp, 2019). The users are predominantly young, with 68% being between 18 and 34 years old (Kemp, 2019). Consequently, SM users are not representative of the population as a whole. However, they still constitute a userbase of large size whose data can possibly complement existing data or even give new insights into migration.

In this study, Twitter was used as a data source for the following reasons:

(1) *Joint geolocation and text data*: Tweets are text data that are also sometimes tagged with locations (place geolocation) or geocoordinates (precise geolocation). Geotags link the physical space with the virtual space of online conversations.
(2) *Rich metadata*: Besides geolocation, tweets are accompanied by a variety of metadata which aid in the interpretation of the data and the detection of automated accounts.
(3) *Accessibility*: Twitter's accessibility has made it the most studied SM platform (Williams et al., 2017). An Application Programming Interface (API) provides access to historical data, although with restrictions to query volume and rates (Twitter Inc., 2022a). For these reasons, Twitter is a preferred data source for spatial analyses, despite estimations that only 1.1% of the adult population can be reached with adverts on Twitter, compared with 18.0% for Facebook (Kemp, 2019).

Figure 1 shows the timeframe for the study, January 2015–December 2018. It was selected to allow for a consistent database with minimal temporal biases, defined by Olteanu et al. (2019) as systematic distortions across user populations or behaviours over time. Preliminary experiments suggested that the geotagging of places became widely used from mid-2014 onwards (Figure 1a). In 2019, Twitter made changes to the geotagging functionality (Kruspe et al., 2021) (Figure 1b). The use of Twitter by Nigerians spiked in 2020, fuelled by the #endsars protests (Ojedokun et al., 2021) (Figure 1c), while in 2021 the Nigerian government enacted a temporary ban on Twitter (Princewill et al., 2021) (Figure 1d). To prevent these events from influencing the study,

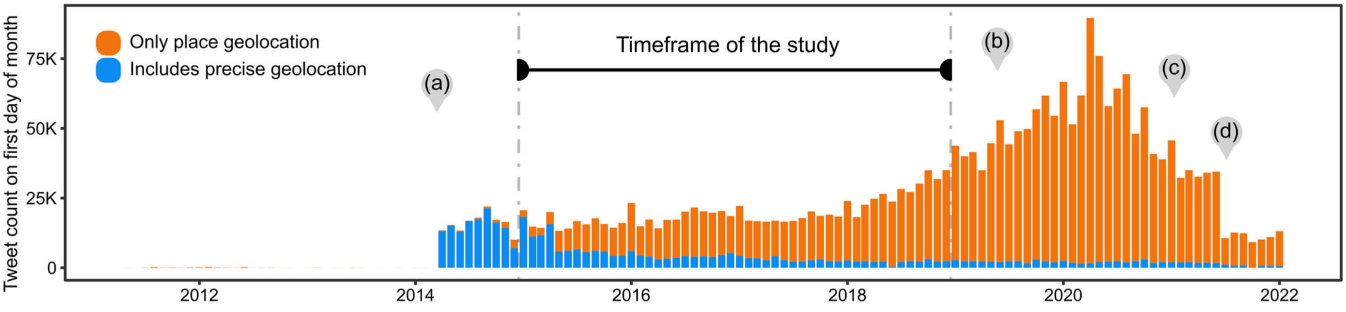

FIGURE 1 Sampled count of geolocated tweets from Nigeria between 2011 and 2022, by geolocation type. (a) First appearance of substantial numbers of geotagged Tweets in Nigeria; (b) change to Twitter geolocation functionality (Kruspe et al., 2021); (c) #endsars protests in which Twitter played a significant role (Ojedokun et al., 2021); (d) temporary ban of Twitter by the Nigerian government (Princewill et al., 2021).

the 4-year period 2015–2018, for which a stable level of general activity and geotagging activity can be found, was selected as the study's timeframe.

As an ancillary data source, data from Nairaland.com, a Nigerian web forum was used. For a description of Nairaland (NL) we refer to Supporting Information Appendix B.

3.2 | Methods

The methodology follows a complementary workflow of spatial analysis and text analysis. On the basis of a source set of Nigerian tweets (S_{NGA}), which is expanded in geographical scope to S_{GEO}, the spatial analysis identifies several groups of users with distinct mobility characteristics (U_0, U_{30} and U_{90} in Figure 2). For these mobile people, a new set of tweets (S_{MPL}) is acquired to analyze their topic interests. Figure 2 illustrates how outputs of the spatial analysis are being used to guide the data acquisition of the text analysis.

3.2.1 | Spatial analysis

For the timeframe, all geolocated tweets from Nigeria were queried using the Twitter API. From the over 28.5 M tweets collected (data set S_{NGA}, Table 1), unique users were identified with the aim to analyze their movement. To exclude automated accounts, a series of filters was applied following previous studies on trajectories by Hübl et al. (2017) and Petutschnig et al. (2020): First, accounts that frequently (for more than 25% of tweets) exceeded the speed of 150 km/h between two sequential tweet locations were excluded. Besides filtering automated accounts, this speed-based filter excluded users who use the geotagging feature consistently in other ways than to refer to their own location (e.g., referring to a place they intend to visit, or the location of an event they comment on). Second, accounts that post more than 150 geolocated tweets in a single day (Hübl et al., 2017), or more than 15 on average per day (Petutschnig et al., 2020), were excluded. As a final measure to improve consistency, only users who were created before the start of the timeframe were selected. For these remaining 116,670 users (U_{NGA}), all tweets from outside Nigeria for the timeframe were acquired to complete their timelines of geolocated tweets (data set S_{GEO}, Table 1).

For the mobility analysis, as small-scale mobility was not of interest in the study, the tweets from S_{GEO} were spatially aggregated to spatial units (SUs) based on the centroid of the place tagged in the tweet. This further serves to anonymize the data. To ensure that the size of the SU did not determine the results, a sensitivity analysis was performed and confirmed that the results were robust to the size of SU influence on the result excepting extreme values. SU of a 30-km radius was used for the remainder of the study. For more information on how the SU was derived, see Supporting Information Appendix A.

The geolocated timelines S_{GEO} contain location histories for 116,670 users. To identify migration events within these histories,

TABLE 1 Twitter data sets used in the study.

Data set	Number of tweets	Number of users	Geolocation	Purpose
S_{NGA}	27,818,148	325,061	Complete	Identification of users U_{NGA}
S_{GEO}	25,021,363	114,278	Complete	Mobility analysis of users U_{NGA}
S_{MPL}	2,629,812	9301	Partly	Text analysis of users U_0, U_{30} and U_{90}

Note: S_{GEO}, global data set for selected users U_{NGA}; S_{MPL}, sampled data set for analysis of mobile populations; S_{NGA}, initial source data set of Nigeria.

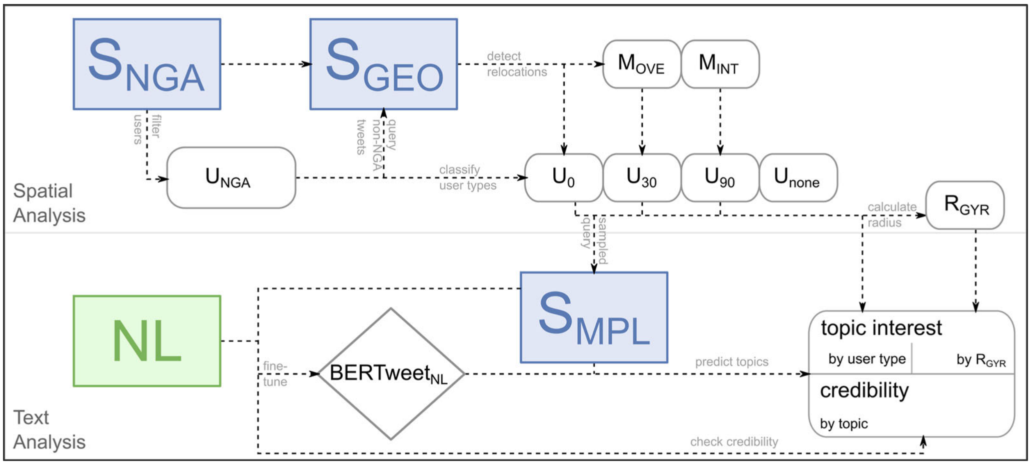

FIGURE 2 Workflow of the study. (Top) Spatial analysis. (Bottom) Text analysis. S_{NGA}, S_{GEO} and S_{MPL} are Twitter data sets. NL is a Nairaland data set (see Section 3.2.2). M_{OVE} and M_{INT} are movements. U_{NGA} is a set of users, of which U_0, U_{30}, U_{90} and U_{none} are subsets (groups) of users. R_{GYR} is the radius of gyration.

the approach developed by Chi et al. (2020) was applied. It can be flexibly tuned to detect migration in various forms and has been calibrated and validated on several data sets, Twitter among them. In this three-step approach: (1) contiguous segments are identified, (2) consecutive segments in the same location are merged and (3) overlaps are removed. Compared with Chi et al. (2020), one of several overlapping segments was allowed to persist if they contained more than 70% of all tweets in the overlapping period. The results are continuous, nonoverlapping segments that indicate a user's presence at a certain location over a certain time (Figure 3).

Segments qualified as residences if their length exceeded a minimum time (*ResLength*). Minimum residency length is part of many definitions of migration but no single definition is universally accepted (Kirchberger, 2021). To demonstrate the flexibility of the approach with regard to migration concepts, two types of movement and four groups of users were distinguished (see Figure 2): The first type of movement was defined as international migration (M_{INT}), which occurred between two residences of at least 90 days *ResLength*, only one of which was within Nigeria. The 90-day window corresponds to the commonly used distinction between visitors and longer terms stays as codified in visa-waiver programmes (Armstrong et al., 2021). It corresponds to the UN definitions of "short-term migrant" and "long-term migrant" which refer to residence periods of 3–12 months, and longer than 12 months, respectively (United Nations Department of Economic and Social Affairs, 1998). The second type of movement was defined as overall mobility (M_{OVE}), which occurred between any two different residences of 30 days *ResLength*. It was intended to capture a wider range of movements that could also include short-term mobility.

On the basis of these two types of movements, four groups of users were distinguished: international, mobile, stationary and others. First, international migrants (U_{90}) were users who displayed at least

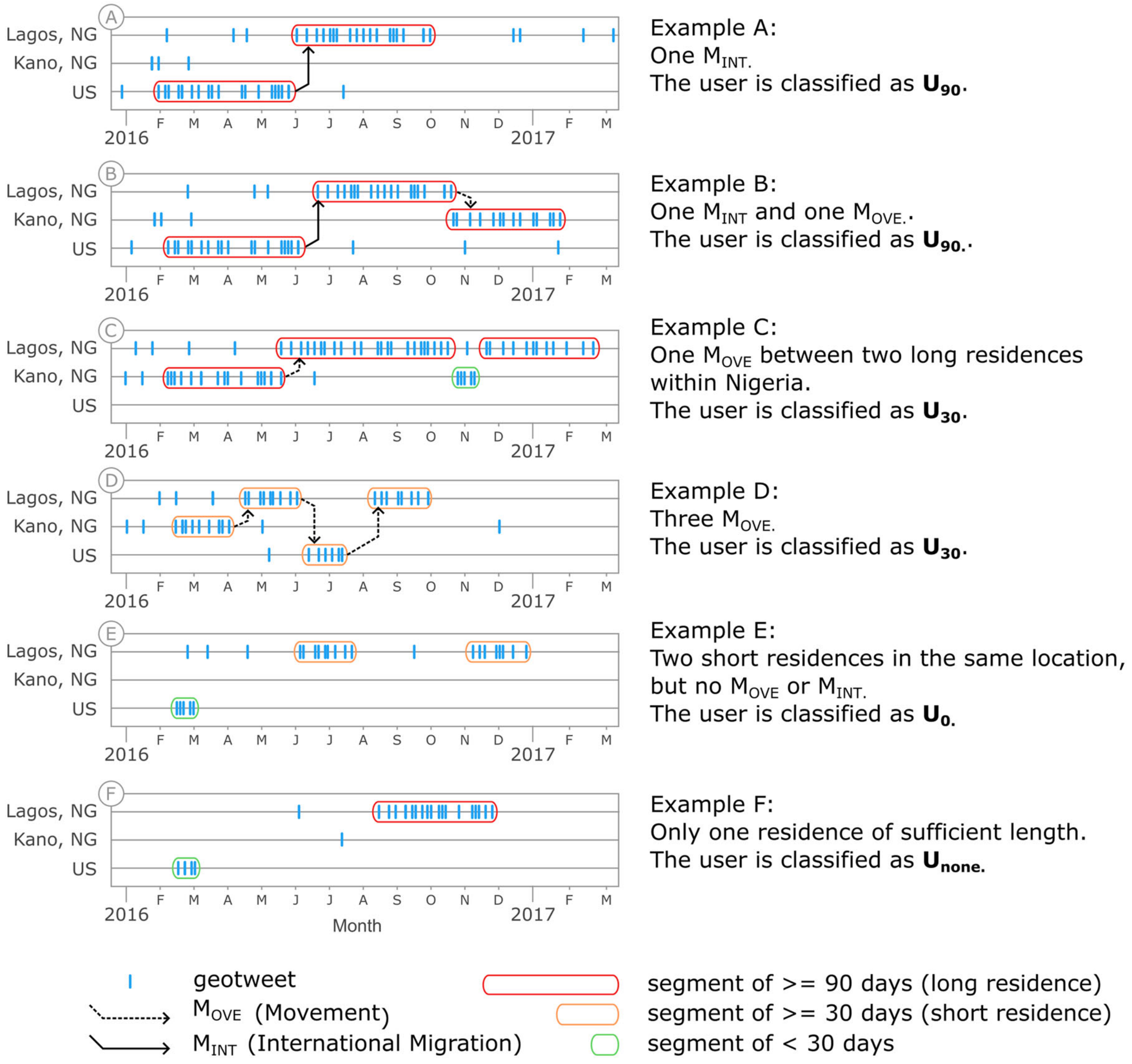
A
Lagos, NG
Kano, NG
US
F M A M J J A S O N D F M
2016 2017
Example A:
One M_INT.
The user is classified as U_90.
B
Lagos, NG
Kano, NG
US
2016 2017
Example B:
One M_INT and one M_OVE..
The user is classified as U_90..
C
Lagos, NG
Kano, NG
US
2016 2017
Example C:
One M_OVE between two long residences within Nigeria.
The user is classified as U_30.
D
Lagos, NG
Kano, NG
US
2016 2017
Example D:
Three M_OVE.
The user is classified as U_30.
E
Lagos, NG
Kano, NG
US
2016 2017
Example E:
Two short residences in the same location, but no M_OVE or M_INT.
The user is classified as U_0.
F
Lagos, NG
Kano, NG
US
2016 2017
Month
Example F:
Only one residence of sufficient length.
The user is classified as U_none.
geotweet
M_OVE (Movement)
M_INT (International Migration)
segment of >= 90 days (long residence)
segment of >= 30 days (short residence)
segment of < 30 days

FIGURE 3 Illustration of the migration detection method on six fictitious user timelines. Consecutive tweets of sufficient density get joined to segments. The succession of segments at different locations is counted as a movement or migration. According to the migrations, users are categorized as either U_0, U_{30}, U_{90} or U_{none}. NG, Nigeria; US, the United States.

one M_{INT}. Second, mobile users (U_{30}) displayed at least one M_{OVE} within Nigeria and no M_{INT}. Third, stationary users (U_0) were those which had several detected residence periods of at least 30 days all at the same location in Nigeria. To prevent inactivity being mistaken for *spatial continuity* (Schewel, 2020), stationary users had to possess at least two residence periods in the same Nigerian SU (cf. Figure 3e,f), and further fulfil the requirement of having a higher number of total tweets than the lowest number of tweets exhibited by any U_{90} or U_{30} (19 tweets). Lastly, other users U_{none} fulfilled none of the other groups' criteria and were not of interest to this study.

As a final analysis step, the movements from one residence to another were aggregated to mobility flows between the SU.

Assigning users and movements to discrete categories allows testing for a priori definitions of migration, but it is not the only way to make mobility tangible. Location histories alternatively allow for the quantification of user mobility on a continuous scale. One possible measure is the radius of gyration (R_{GYR}), the mean distance of a user's geolocated tweets from their collective centroid (see Zagheni et al., 2014 for an example). To test this alternative method of capturing mobility, R_{GYR} was calculated for U_0, U_{30} and U_{90}.

3.2.2 | Text analysis

The objective of the study was to identify whether the previously identified differences in mobility also extend to differences in topic interests of the particular user groups, expressed by the tweeting about certain topics.

The geolocated tweet timelines S_{GEO}, used in Section 3.2.1 to detect mobility, were not suitable for the analysis of topics, as they only contained geolocated tweets. These are likely biased towards certain topics that have a strong spatial component (e.g., *travel*) or ties to particular locations (e.g., *sports*). To overcome this limitation, a new set of tweets S_{MPL} was queried without a requirement for geolocation.

A technique of random sampling, stratified by users, was applied to download a representative selection of tweets from all users who were in the previous step classified as either international, mobile or stationary. This sampling was necessary for a couple of reasons: (1) Acquiring the users' complete tweet histories for the time between 2015 and 2019 is theoretically possible, but practically unfeasible due to the high number of tweets. And (2), a complete data set would be highly biased towards more active users (Li et al., 2013; Zagheni et al., 2014).

Random sampling of tweets from users' tweet histories is not offered by the Twitter API. As a heuristic alternative, a pseudo-random stratified sampling approach was implemented by sampling 48 randomly spaced 7-day intervals throughout the timeframe, and acquiring for every user up to 10 tweets from within each of these intervals. Tweets from highly topic-specific sources or automated platforms were discarded.

The resulting data set S_{MPL} comprises 2,731,483 tweets from 9672 users. Basic summary statistics across user groups are provided in Table 2.

For the classification of topics within the tweets, we apply a machine-learning model. To train this model, labelled training data are necessary. As an alternative to manual labelling of tweets, implicitly labelled training data from the Nigerian web forum NL were acquired. NL consists of many subforums dedicated to certain topics, like, *politics*, *sports* and *travel*. By choosing the appropriate subforum for their comments, users are implicitly labelling their own texts. Therefore, such hierarchically structured web forums can be seen as a labelling environment in which labels are offered top-down and assigned by the web community bottom-up, with certain community members having more control over the labelling than the majority. Compared with a manual labelling of posts or definition of topic-specific keywords by experts, automatically learning the labels from the online community is more likely to reflect the themes discussed in online spaces and the language that is being used to discuss them. While this approach does not avoid subjectivity or normativity, it makes for a better fit to the microblog format. In total, 2,091,491 NL comments spread over 40 topics were acquired using a web-scraping approach. Of the NL comments, 80% were used for training, 10% for validation and 10% for testing the model and calculating the topic-specific accuracies for the plausibility check. For a more detailed description of the NL data set, see Supporting Information Appendix B.

To analyze the interests of Twitter users, a state-of-the-art language model was trained on the NL data set and then used to assign topic labels to all the tweets in S_{MPL}. The feasibility of a similar approach has been demonstrated by Fiallos and Jimenes (2019) who labelled the interests of Twitter users' based on a classifier trained on Reddit comments.

BERTweet (Nguyen et al., 2020) was used as the backbone of the model. It is a language model based on the transformer architecture (Vaswani et al., 2017) that is pretrained on a large English Twitter corpus. This backbone converts the text into vectorized embeddings

TABLE 2 Statistics of users over S_{MPL}.

		Number of tweets				R_{GYR} (km)			
	Number of users	Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD
Stationary U_0	3748	13	480	282.4	128.7	0.6	5879.7	413.2	714.7
Mobile U_{30}	4655	1	480	276.4	128.0	1.5	8204.4	864.4	1258.9
International U_{90}	898	22	480	316.9	117.9	66.4	10,045.5	2478.4	1463.7

Note: R_{GYR}, radius of gyration; SD, standard deviation; U_0, stationary users; U_{30}, mobile users; U_{90}, international migrants.

which feed into a fully connected classification layer. The entire network, including the transformer backbone, was then fine-tuned to the task of predicting the topic for each NL comment. Afterwards, the trained classification model was applied to the Twitter data set S_{MPL} to predict the topics for each tweet.

This approach is one of unsupervised domain adaptation (see Ramponi & Plank, 2020, for an overview) in which S_{MPL} constitutes the target domain (D_T) and NL constitutes the source domain (D_S). Despite many similarities between NL and the Nigerian Twittersphere, there remains a domain shift between D_T and D_S due to differences in userbase, moderation or user options. To reduce this shift, tweets and NL comments were preprocessed identically, removing obviously platform-specific features (such as the retweet marker "RT") and masking unique items, such as URLs or phone numbers. As in the original training of BERTweet, comments shorter than 10 tokens were excluded. On grounds of representativeness, the small proportion of non-English tweets and NL comments were not excluded, although their classification likely added challenge of the task. To the model, two further adjustments were made compared with the reference implementation. First, to account for unequal representation of topics in our NL data set, cross-entropy loss with class weights was used (Torch Contributors, 2019). Second, a domain-adversarial domain adaptation approach was applied during training (see Ganin et al., 2016) to promote the learning of domain-invariant features. It was also found to serve as a regularization technique that reduces overfitting and improves accuracy.

The model was trained for five epochs and reached a validation accuracy of 42.5% on withheld NL validation data, an acceptable result considering multiclass prediction on 40 classes is a much more difficult task than single-label prediction (Quercia et al., 2012). The accuracy is comparable to the 42.1% accuracy of the best-performing model in the TREC-IS 2018 challenge involving 25 classes of disaster-related tweets (McCreadie et al., 2019).

Within each predicted topic, the significance of differences between the groups U_0, U_{30} and U_{90} was determined via the Kruskal–Wallis rank sum test.

Due to the lack of labelled tweets, the model's performance on tweets cannot be gauged with traditional validation metrics. As an alternative, a threefold credibility check was applied to each topic.

First, the topic-specific F1-score was calculated on the predictions on a withheld NL test set that comprised 10% of all NL comments. If a topic could be accurately classified in D_S, the same would be true in the D_T. The F1-score was calculated as follows:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1score = \frac{2 \times Precision \times Recall}{Precison + Recall},$$

where TP is the true positives, FP the false positives and FN the false negatives.

Second, characteristic weekday activity patterns were checked for similarity. When the topic exhibited a similar pattern in the tweets as in the NL data set, as quantified by the Pearson correlation coefficient (COR) between their activity shares over weekdays (Leydesdorff, 2005), the model was able to recognize a particular topic in the Twitter domain. This check was not applied for topics where a coefficient of variation (CV) lower than 0.05 indicated the absence of any noticeable weekly pattern (e.g., *gaming* in Figure 4).

Third, ancillary data sets of NL comments NL_{NEW} and of tweets S_{NEW} were acquired from a more recent period (2021–2022). From NL_{NEW} keywords were extracted that did not appear in the older NL data set. These keywords were assumed to represent novel concepts (e.g., "covid19"). If they appeared in similar topics in S_{NEW} as in NL_{NEW}, it supported the assumption that the Twitter topics matched NL topics. To quantify this similarity for each topic, the cosine similarity (Egghe & Leydesdorff, 2009) between NL_{NEW} and S_{NEW} was calculated using the topic's shares of the usage of the aforementioned keywords as features.

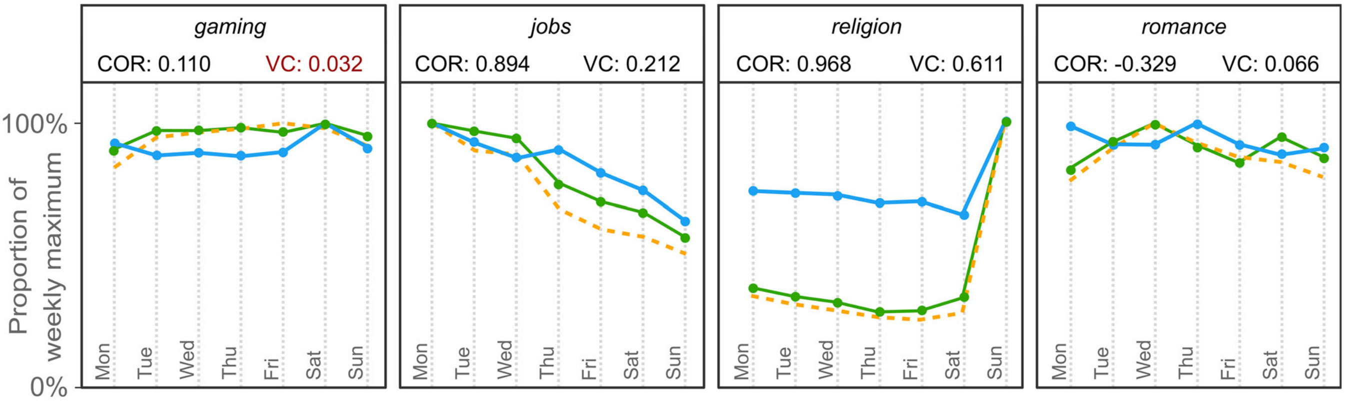

FIGURE 4 Weekly patterns and correlations of example topics. (Green) True labels of all NL posts. (Orange) Predictions on NL test data set. (Blue) Predictions on Twitter data set. *y*-Axis: Proportion of the weekly maximum by topic and data set. COR, Pearson correlation between NL and Twitter patterns; CV, coefficient of variation; NL, Nairaland.

To condense this three-pronged check into a tangible measure, the predicted topics were grouped into four discrete grades of credibility from A (most credible) to D (least credible), based on whether they failed to meet a grade-specific threshold in all three checks. Thresholds of 0.5, 0.4, 0.3 and −1 were used for grades A, B, C and D, respectively.

To complement the quantitative metrics of plausibility, we perform a qualitative reading of sampled tweets on those topics for which we find the strongest differences between user groups, as well as the topic which likely holds the most migration-related information. Our goal is to better understand how much variation exists in the content that is algorithmically bundled into a topic. For each topic, 100 tweets were sampled for each user type U_0, U_{30} and U_{90}, for a total of 900 tweets across three topics. Each tweet was assigned a code, based only on its raw text and the content of any linked web pages or images. Following grounded theory methodology (Charmaz, 2006), codes were not defined by preconceived hypotheses but developed from the observations in the data in an exploratory manner, although with a focus on mobility-related information. Consequently, the goal was not an overarching framework of themes, emotions or styles across all topics but a specific set of codes for each topic which best represents the variation within it.

4 | RESULTS

4.1 | Results of the spatial analysis

Figure 5 displays the migration flows between the most connected places for overall movements (M_{OVE}) and international migrations (M_{INT}). Of 10,101 M_{OVE}, 50.9% were within Nigeria, 18.8% were outgoing, 18.4% were incoming and 11.4% were between countries outside Nigeria.

Of the 1057 M_{INT}, 48.0% were incoming to Nigeria while 52.0% were outgoing. The capital Abuja and the largest city Lagos are measured with the strongest international ties, most connecting to the United States and the United Kingdom.

As for the users, the number of mobile users (U_{30}) was 4794, of which 3229 (67.4%) moved purely within Nigeria. For roughly a third (35.3%) of the U_{30}, we observed multiple M_{OVE} in the timespan.

The international class U_{90} consisted of 926 users. For 111 (12.0%) of these, we observed multiple international migrations in the timespan.

This means that we measured some mobility for 5720 (4.9%) of our 116,670 analyzed users.

Of 3857 stationary users (U_0), most resided in Lagos (49.2%), Abuja (20.0%), Kaduna (3.5%), Ibadan (3.3%) and Port Harcourt (3.2%).

The gyration radius is right skewed with a median of 223 km and a mean of 840 km. Interestingly, the distribution of its decadic logarithm is clearly bimodal, with two peaks corresponding to radii of approximately 200 and 2200 km (Figure 6).

4.2 | Results of the text analysis

The credibility of the domain adaptation varied strongly by topic. In this section, we provide results for credibility grades A, B and C. The complete results of the credibility check and the topic analysis for all topics are provided in Appendices C and D.

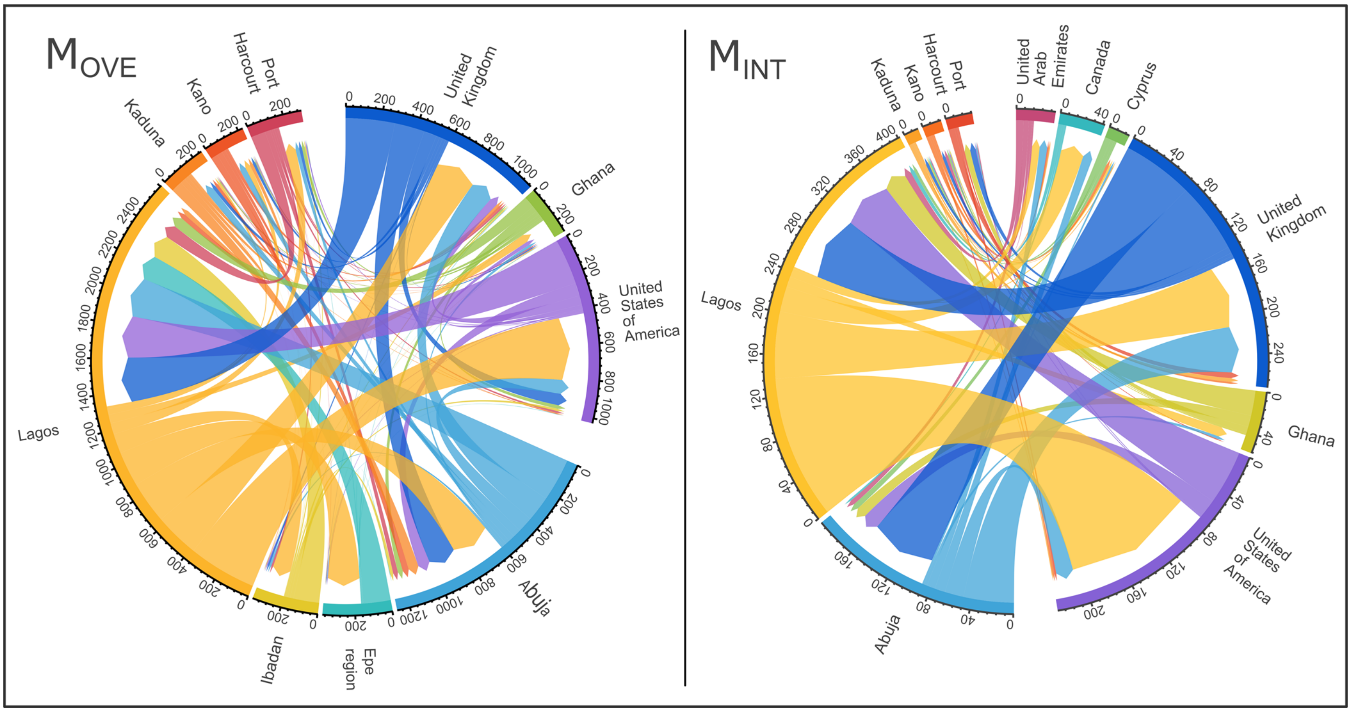

FIGURE 5 (Left) M_{OVE}s between SU with at least 250 M_{OVE}s. (Right) M_{INT}s between SU with at least 25 M_{INT}s. SU, spatial unit.

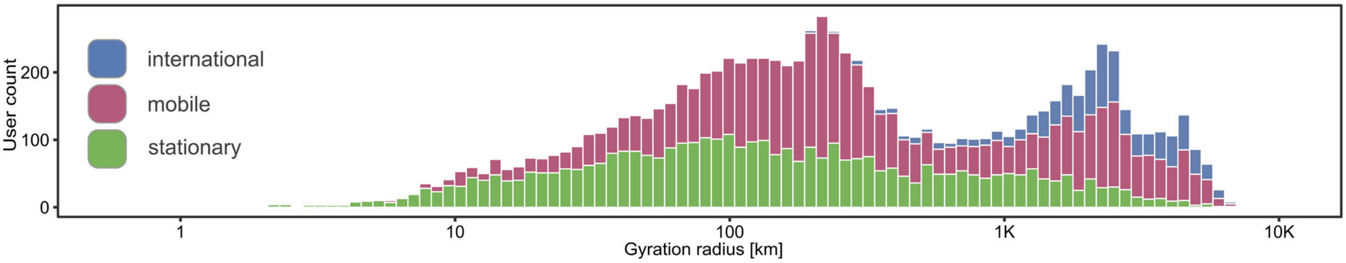

FIGURE 6 Histogram of the decadic logarithm of gyration radii.

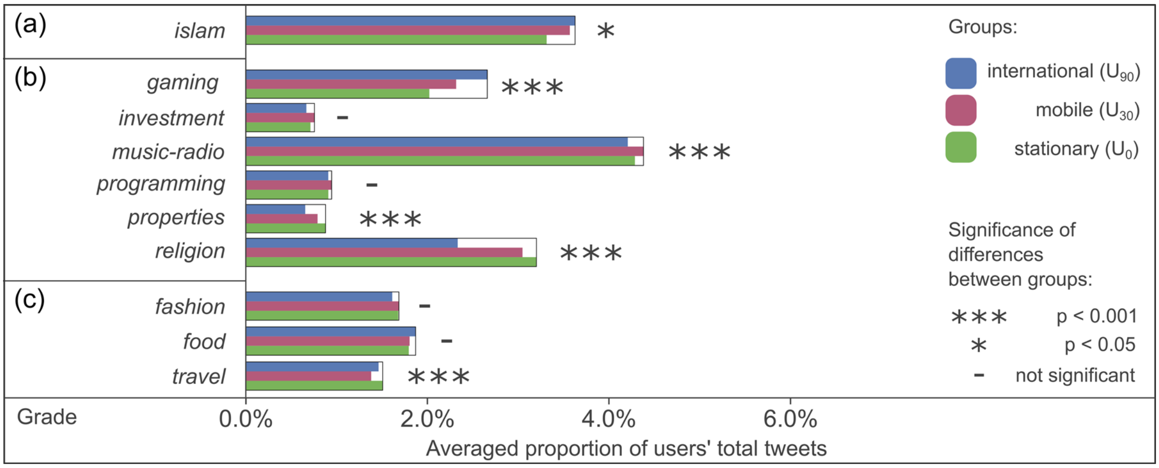

FIGURE 7 Averaged topic interest of various user groups. Topics are graded by credibility into grades (a), (b) and (c). Significance of differences between groups was determined via Kruskal–Wallis rank sum test.

Figure 7 displays for each group of users the average share of users' tweets. Striking differences between user groups could be seen in the topics *islam*, *religion* and *gaming*.

Islam has been assigned to 3.86% of U_{90} tweets, but only 3.68% of U_{30} and 3.31% of U_0 tweets. The reverse pattern was found for *religion*, with 2.64%, 3.44% and 3.77% for U_{90}, U_{30} and U_0, respectively. *Gaming* made up 3.05% of U_{90} tweets, but only 2.54% of U_{30} and 2.11% of U_0 tweets. The topic *properties* seemed to be of interest for U_0 while *music-radio* was of interest to U_{30}. Notably, U_{30}, intended to represent mobile users, had lower shares in *travel* than either U_{90} and $U_{0.}$

For the topics *gaming* and *religion*, the strongest differences in interest were found across user groups. They, along with topic *travel*, were qualitatively evaluated. The codes, displayed in Figure 8 and described in Supporting Information Appendix C, give indication about the content of the topics. *Gaming*, more popular with international migrants, was found to include a mix of entertainment-related subtopics, most commonly football, but also other sports, games and music. Public conversation about aspects of public life (celebrities, fashion), private life (relationships and pets) and politics were also common, making up around 45% of the sampled tweets. *Religion*, on the other hand, was thematically homogeneous, with around 18% of the tweets not clearly related to religion. It was found that the religious discourse took a variety of forms, ranging from bible quotes to proclamations of devotion or prayers to discussions about matters of politics or proper behaviour. Dominant in the topic *travel* was various forms of mobility-related information at various spatial scales (from local traffic updates by users to international news by journalists) as well as news, information and discussions about places or events with an explicit spatial location. Comments on the state of transport infrastructure and transport service providers were predominantly critical. Intention to travel or migrate was only occasionally (2%) expressed, while sharing of personal travel experiences was more common (~13%). Around 25% of tweets were not clearly mobility related.

As described in Section 3.2, the gyration radius R_{GYR} provides a continuous measure of mobility, which allows the exploration of the relationship between mobility and interest without grouping users. Due to the large number ($n = 9672$) of users, locally estimated scatterplot smoothing (LOESS, Cleveland & Loader, 1996) was applied to aid exploration of patterns in the relationship from a large number of tweets (Figure 9). These patterns suggested that increased mobility might correlate with lower interest in religion, as was previously observed in the groupwise topic interest (Figure 7). The inverse was true for *gaming* and particularly *travel*, where, past a certain point, the interest in the topic seemed to be logarithmically related to R_{GYR}. For *islam* and *properties*, the curves suggested a more complex and multimodal relationship.

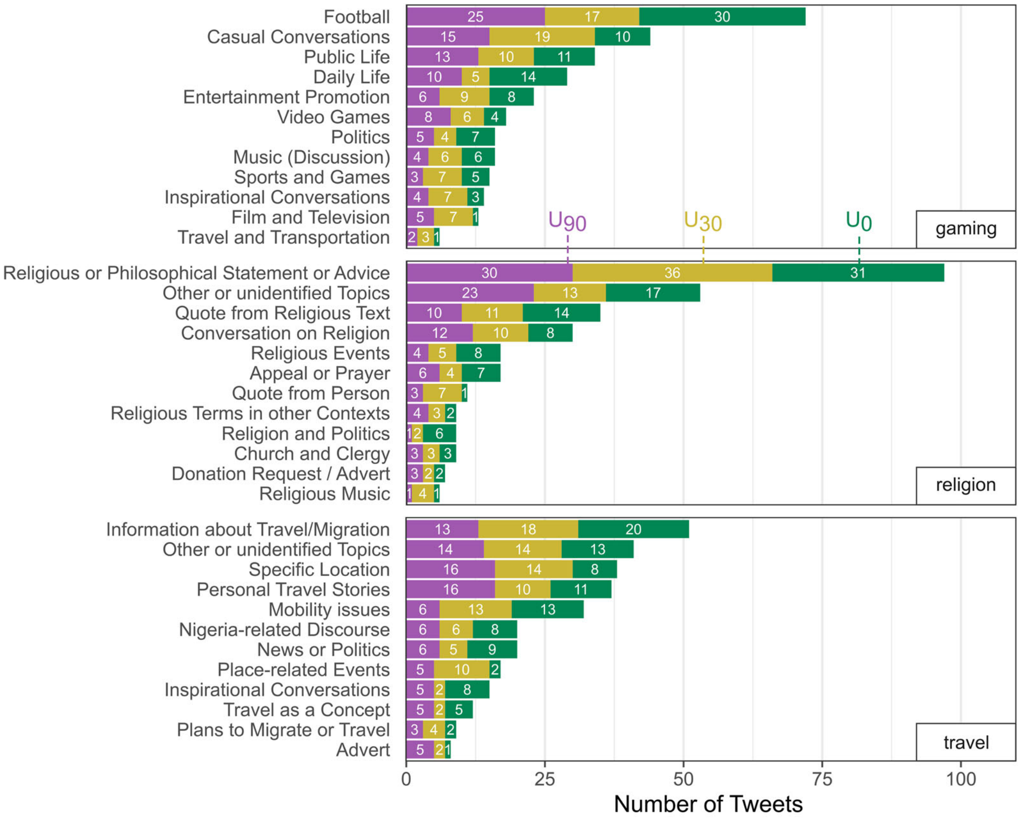

FIGURE 8 Frequency of codes in the qualitative reading of topics gaming, religion and travel. For each topic, 100 tweets were randomly sampled per user group (U_0, U_{30} and U_{90}). Numbers on the bars refer to the tweet count.

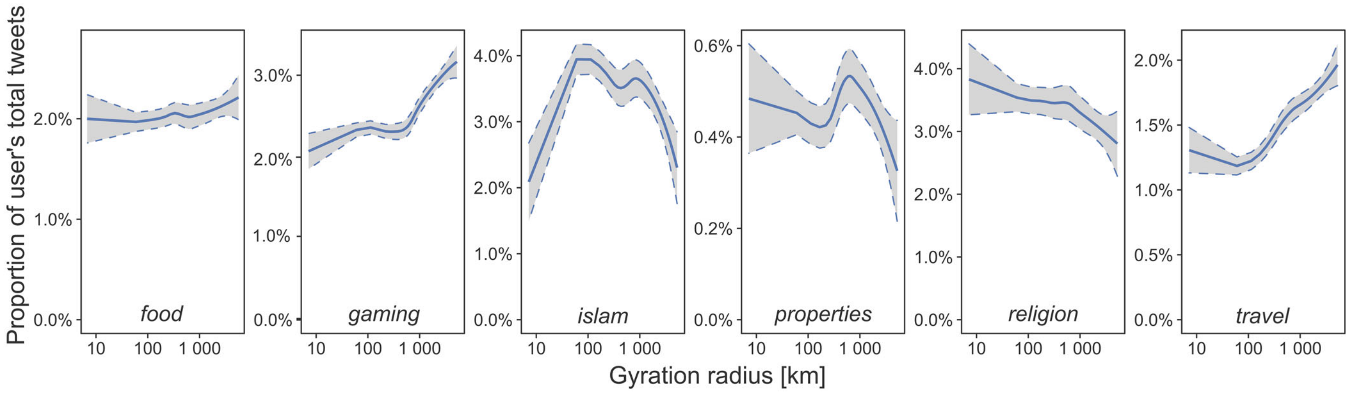

FIGURE 9 Locally estimated scatterplot smoothing curve of topic interest versus the decadic logarithm of user gyration radius. Due to the large number of points (9672) only the fitted curves are shown. Note that the *x*-axis is logarithmic. The shaded areas between the dashed lines indicate the span of the 95% confidence interval.

5 | DISCUSSION

5.1 | Key findings of the case study

The detected mobility flows suggest that most international migrations of Nigerian Twitter users connect to large urban centres and support the findings by Kirwin and Anderson (2018) that among Nigerian urban residents, Lagosians are more likely to want to migrate abroad. Besides Lagos, the smaller Abuja has almost comparable international ties, likely due to its unique status as the state's capital. We found the strongest international connections to the United States and the United Kingdom, which agrees with polls conducted at the start of our study period (NOI Polls Ltd., 2015). However, the well documented movements of refugees and internally displaced persons (IDPs) in the north of Nigeria (International Organization for Migration [IOM] & Displacement Tracking Matrix [DTM], 2022) were not reflected in our findings, confirming that refugees and IDPs are likely less or not active on Twitter, as discussed by Petutschnig et al. (2020).

Our results suggest that international migrants have, on average, a higher interest in the topic *islam* than stationary users, who in turn seem to have more interest on the topic *religion*. This is in contrast with previous work by Bloch et al. (2015) and a survey conducted by Kirwin and Anderson (2018), which found that within Nigeria, Muslims are less likely to want to migrate abroad. It is possible that the relationship between religion, *islam* and mobility is more complex than was previously assumed, at least for the demographic of Twitter users (see Section 5.3). This view is also supported by the LOESS regression for the topic *islam*, which painted a more complex picture than the categorical results suggested. International migrants displayed significantly higher interest in the topic *gaming*, which consisted of a surprising variety of entertainment-related subtopics, such as music, celebrities and above all football. These may not, by themselves, be significant migration factors. But football is of great importance to Nigerian culture, and the attachment to European football brands creates strong international ties with consequences both positive and troubling (Igwe et al., 2021). The topic *travel*, selected because its label suggests relevance to mobility studies, curiously was observed with the highest popularity with stationary users, albeit by only a small margin. As we found in our qualitative reading, the topic comprises a large variety of mobility-related themes, conversations and information, from a local to global scope. Stationary and mobile users alike frequently commented on inadequate transport infrastructure, traffic obstructions or congestion, which are serious problems in Nigerian cities (Afolabi et al., 2017) and have a wide range of negative impacts on people and the economy (Economic Intelligence Unit, M. of E. P. & B., 2013). For international migrants, on the other hand, we found that reports of their own journeys and specific locations prevail. While these insights are based on a relatively small sample of 900 tweets, they hint at the type of insight that can be gained about the differences and similarities between international migrants and stationary people using such an approach as presented in this paper. In conjunction with our quantitative results, we found indications that mobility is of importance even for people who would be, at a large scale, considered stationary. At a methodological level, we found that our approach to identifying general topics accomplishes its inclusivity at the cost of masking the complexity of information contained within. This complexity can appear as diversity in subtopics, as in the example of *gaming*. Or, as in *religion*, as a variety of conversation types (e.g., advice, discourse, debate, prayer, quote). Or as in *travel*, the information can vary in spatial and temporal scope, from ad-hoc traffic updates by local users to professional news reporting about developments in international politics. Therefore, statistical analysis of any topic should always be accompanied by in-depth reading of at least a sample of tweets. Future studies could aim at digging deeper via a more detailed inspection of tweet messages in a conversational context, considering the line of argumentation in the tweets, and including statistics about linguistic features and metadata in a multimodal approach.

5.2 | Policy implications

We find that our results are plausible and conclude that with SM data and our methodological approach, it becomes possible to capture issues that preoccupy mobile users at a general level. Thus, we offer means to approximate the migrants' perspective, and the data allow a detailed evaluation of the textual content based on this. Compared with the approach by Kim et al. (2021), the approach presented in this paper is more demanding (in terms of data) but also versatile in the classification of mobility and content. For researchers, this grants flexibility in the criteria they use to identify migrants, enables the combination of multiple criteria in an ensemble (as proposed by Johnson et al. (2016), and could contribute to a better understanding of the link between short-term mobility and long-term migration (Mau et al., 2015). Altogether our presented approach is likely less suited to generating and forecasting demographic data and more useful to studies of public opinion and research in the social science domain, where it can serve as a timely intermediate between official statistics and in-depth qualitative research. We believe it will be particularly useful for comparative studies of international migration, where the granularity and extensive coverage of the geolocated SM data provide a clear advantage over traditional data sources (Bosco et al., 2022).

Our approach provides an overview of migrants' general interests but naturally is not very detailed on any specific issue. More detailed and qualitative approaches, which focus on specific topics that are clearly defined by the researcher via keywords and hashtags, are going to remain vital. Using different approaches in conjunction, we believe, is a way by which researchers can gain a more complete and less biased picture of migrants.

However, understanding what occupies migrants is not merely of academic interest, but beneficial to the management of migration and its impacts. As Kim et al. (2022) show, topics can be the key to understanding attachment to places of origin or destination. And this can result in substantial economic impact: In Nigeria, remittances, given by ethnic Nigerians in diaspora, constitute a significant source of foreign currency with a $20.9 billion inflow in 2022 (World Bank, 2022). Neither this inflow nor its benefits should be taken for granted, however, as Didia and Tahir (2022) found remittances do not automatically enhance economic growth, with a large portion of remittances going towards consumption and social insurance rather than investment. They recommend that by better understanding the diaspora and earning their trust, remittance behaviour could be improved and channelled towards economic growth. Knowledge from SM could support this by informing initiatives like the Nigerians in Diaspora Commission (NiDCOM, 2021) about ways to connect with the diaspora and learn about topics in which they might be interested to invest in. Our findings indicate that compared with stationary Nigerians, internationally mobile Nigerians engage more in the public discourse on sports and other entertainment while engaging less in discussions about—or public displays of—faith. Engaging the diaspora on topics that are suited for online discussion can be an important step to connect with the users on issues that matter to them, and possibly encourage them to invest remittances on projects related to these issues. On the other end of the mobility spectrum, those SM users who would prefer to remain in their communities can also benefit from having their collective voices heard and understood. Policy makers can engage and support them in issues that matter to them, in the long run, also increasing the resilience of the communities of stationary Nigerians.

Beyond the tangible benefits, we hope that the study of migrants' interactions on everyday issues can preserve the human perspective in the migration discourse and help migrants be seen as more than mere atomic parts of a highly political issue. In our case study we discovered a personal side of migrants, whose interactions are also on nonpolitical topics, such as relationships, sports or films. We hope that this can help build empathy and understanding, which will indirectly benefit all aspects of migration management.

5.3 | Limitations and ethical considerations

Despite the proven capability of the presented approach, we acknowledge several limitations. We demonstrated that different forms of mobility can be identified within the same data set by varying the parameters of the algorithm. However, long information gaps between identified residence segments, resulting from irregular tweet activity of users, make it practically impossible to precisely define the residence length and the number of relocations for most users. The identification of seasonal migration patterns is greatly hampered by this data irregularity. Simpler migration patterns are less reliant on the continuity of tweet timelines and can be confirmed or rejected for more users. Consequently, while the distinction of mobility at different rates (e.g., weekly, monthly, seasonal) is only feasible for few very active users, the differentiation between mobile and stationary users can be made with a fair degree of confidence for most users.

Similar considerations apply in the spatial dimension. A practical constraint to spatial granularity is posed by the relatively small number of tweets with precision at a point or neighbourhood level (Kruspe et al., 2021). At the city scale, however, we can confirm the robustness of our approach towards the size of the SU with a sensitivity check (see Supporting Information Appendix A). We conclude that the data are well suited for migration studies at intercity, national and international scales and over long timespans.

The LOESS regression of mobility versus topic interest produces intriguing patterns that paint, for some topics, a different picture than the results by groups. While the method is exploratory, it illustrates that, unlike traditional census or survey data, geolocated tweets offer the opportunity to capture mobility without a priori definitions of migration. Another promising use of continuous mobility information is the identification of returners, which Pappalardo et al. (2015) identified as people for whom recurrent movement constitutes a large part of their mobility and are clearly distinct from explorers who spread their movement over a larger number of locations.

Altogether, we find that geolocated SM data support studies using a wide range of mobility concepts and can also theoretically be implemented in a large variety of settings. However, there are practical limitations: Changes in Twitter's policy, changes in user activity and governmental restrictions affect the stability of the data basis over time. Across space, the availability of geolocated tweets is likewise not homogeneous. The population's affinity for geotagging, a requirement for the analysis of mobility, varies across user types and across countries (Huang & Carley, 2019), as does mobile phone penetration (Gollin et al., 2021). Thus, while our approach theoretically supports many settings, we highly recommend that preliminary experiments confirm the existence of a suitable data basis for each particular application. More than a matter of technical viability, this is a matter of representativeness.

It is generally agreed that Twitter users are not a representative sample of the population (Spyratos et al., 2018; Taubenböck et al., 2018). Despite the difficulty in assessing specific biases (Wang et al., 2019), it is commonly asserted that SM platforms cater especially to a young and urban population (Gollin et al., 2021; Hughes et al., 2016). Malik et al. (2015) found for the US that a bias in geotag usage further limits the representativeness of findings. This limitation naturally extends to trajectory-based analyses: According to Armstrong et al. (2021) the forms of mobility detected by Canadian tweets correspond less to migrants than to business travellers or transnationals. Our own results support the finding by Petutschnig et al. (2020) that Twitter is ill-suited to inform about rural migrations and refugee movements. On the flipside, it is suitable for the study of transnationals, business travellers and urban–urban migrants—forms of mobility that is widespread, but underrepresented in research (Armstrong et al., 2021). On the basis of continued urbanization and improved access to telecommunication in many countries across the globe, we expect an increase in the number of people which can be mapped with SM approaches.

We now turn from general demographics to a group of accounts that are not controlled by a real person, but by some form of automated algorithm (i.e., bot). The challenge of detecting and filtering bots on Twitter has received considerable attention (Efthimion et al., 2018; Orabi et al., 2020; Subrahmanian et al., 2016). While not all types of automation are malicious (Orabi et al., 2020), confusing bots with actual users will distort and discredit any analysis intended to study human behaviour. In addition to simple initial filters (see Section 3.2), we applied a posterior bot check. We found that both (A) the migration detection algorithm (see Section 3.2) and (B) the selection of general sources (see Section 3.2.2) were effective, albeit imperfect barriers to nonmalicious automation, and conclude that the influence of bots on our study is small.

We addressed the lack of labelled validation data by applying a threefold credibility check which supports the interpretation of the results by quantifying credibility. Of 40 considered topics, 30 do not reach our standards across all checks. Rather than outright dismissal, closer investigation of these topics might lead to further insights. The topic *sports* illustrates this clearly: While the temporal patterns between NL and Twitter show only a weak correlation, we observe a high cosine similarity and classification accuracy. It is certainly possible that this indicates an unsuccessful adaptation of sports from

NL to Twitter. But an alternative, much more interesting explanation is that the mismatch in temporal patterns results from the different ways in which the two media platforms are used to discuss the topic—perhaps Twitter lends itself to exchange about momentary events, like, during a game, while web forums are used for predictions about games or general debates. In-depth analysis of specific (sports) events could shed light on this question.

An important aspect of our approach is that the labels are not predefined by the researcher but rather, in an implicit manner, by the users. We find that this can lead to perceived misalignment between content and label. For instance, after qualitatively analyzing the topic of gaming and finding that it includes sports, music and film in addition to video games and board games, we believe that the label "entertainment" might be more accurate. Changing the name of the label is always an option, of course, but risks imposing the researcher's biases about semantics and related concepts. We refrain from any relabelling in this study because we allow for the possibility that Nigerians use English in different ways. Nevertheless, we believe it is a valid option when used in conjunction with a qualitative reading.

Comparison of the interests between topics is discouraged due to substantial differences between precision and recall. Where the precision is lower than the recall, the number of posts has been overestimated, and vice versa. But, assuming roughly similar error rates between groups of users—something we cannot validate in our current setup—even the relatively low validation accuracy of 42.5% merely creates noise that diminishes differences between user groups, but does not distort them (cf. Figure 3, where the weekly patterns of jobs and religion are weakened, but not completely eliminated in the Twitter data). Not only does this mean that comparisons between groups are still credible but also that differences between groups are likely even stronger in reality than they appear in our results. Many errors are due to confusion between similar and overlapping topics, such as phones and phone-ads. While we assume that these NL subforums are separate for good reason, if a qualitative reading reveals that distinguishing such topics does not provide any relevant insight for the analysis, the topics could be manually merged for ease of interpretation and improved accuracy. For an example of such postprocessing, we refer to Supporting Information Appendix F.

Besides the brevity of texts and the high number of (potentially overlapping) topics, the linguistic diversity is likely a main challenge to accurate prediction. Nigeria is one of the most linguistically diverse places in the world (Orekan, 2010), and several different languages are used on Twitter by stationary users as well as by mobile users. In general, it is to be expected that code-switching and code-mixing on Twitter play a great role in migration contexts when the social networks of the mobile users change, for example, when they meet other migrants during their travels or become acquainted with members of the neighbourhoods they plan to settle in. Theoretically, our approach is language agnostic, but most current state-of-the-art models are pretrained on mostly English corpora, and the predominant language on Twitter itself is English. In contexts where other languages and different varieties of English are used (possibly even within one tweet), the greater linguistic diversity requires language models with a higher capacity and flexibility. But besides this being a technical challenge, there is also an ethical concern. From an ethical standpoint, inequalities in the representation of languages are propagated by the dominance of English training data sets: Because speakers of minority languages are not well represented in the training data, they are not only less acknowledged, but they are also more likely to be misunderstood. In the case of African languages, there are efforts towards better representation in the field of NLP (Masakhane, 2022).

Further, while the use of geospatial SM in approaches such as ours has the potential to generate substantial benefits by enlarging the knowledge for policymaking (see Section 5.2), we must acknowledge it could also be misused in the same fashion.

The data we acquired inform about the affinities and behaviours of human individuals. Such information necessitates ethical considerations of privacy, consent, anonymity and potential harm that could arise from the study (Kochupillai et al., 2022; Townsend & Wallace, 2016). This is particularly the case in the migration context, as migrants can be a particularly vulnerable group (Sîrbu et al., 2021). Point aggregation to large SU, as we applied in our processing, is a way to preserve user's geoprivacy and anonymity (Kounadi & Resch, 2018) and allows for the ethical use of data without the acquisition of consent (Williams et al., 2017). No group was smaller than five users, and SU with less than five detected resident users were excluded to preserve their anonymity. Consequently, we believe that privacy and anonymity of users are not at risk by the presented approach. If follow-up studies more closely inspect users and tweets with qualitative methods, it is vital and feasible that the identity of the users remains undisclosed or cannot be easily retraced. Also, if future developments allow for significantly higher spatial precision, geoprivacy should be incorporated into the study design (see Kounadi & Resch, 2018).

The users' right to be forgotten ought to be respected on a public platform, such as Twitter, by use of the dedicated batch compliance API endpoint (Twitter Inc., 2022b). Over the course of this study, we continuously updated the data and the results, but the present paper can, of course, not be altered in the same fashion. Thus, we only present aggregated results wherein no individual can be identified. We believe that in this manner, the requirements of scientific research and the rights of data subjects are reconciled.

Finally, ethical concerns result from SM data's ease of use. Compared with traditional surveys, whether in the field or online, SM data can inform about people at reduced costs and engagement. It is conceivable that the researchers' understanding of the geographical contexts is likewise reduced, and that engagement with the studied population suffers, while algorithmic biases are propagated. Insights from interviews and field surveys can help identify and mitigate such issues. Thus, we argue that SM data should not replace traditional sources of information but go hand in hand with them in a complementary analysis.

6 | CONCLUSION

In this paper we studied the feasibility of analyzing migrant interests based on geolocated SM data. We found that the data are suitable for the study of urban demographics and at large spatial and long timescales.

In this context, it allows migration researchers to observe—at a hitherto unprecedented scale—the subjective perspectives of people who move, and of those who do not. Flexible in application, our approach can contribute to migration research by filling a gap between qualitative studies and large-scale demographic data, and support the study of migration as a spectrum rather than a condition. In our case study we found that for most topics, users' interest is related to their mobility. It is very likely that the underlying causalities are complex and can be only understood by integrating the data with the knowledge of locals and subject-matter experts. In that light, the joint analysis of geolocation and texts holds a potential that we have barely scratched the surface of. We are looking forward to what future studies in migration research will be able to unearth.

The tweet IDs to reproduce the research can be requested from the author. The full data are not made available due to ethical restrictions and the Twitter Developer Agreement and Policy.

ACKNOWLEDGEMENTS

This study has been conducted as part of the European Union's Caroline Herschel Framework Partnership Agreement on Copernicus User Uptake (FPCUP) under Grant Agreement No. FPA 275/G/GRO/COPE/17/10042, Action 2019-2-11 “Downstream service/application development for monitoring of environmental indicators”, the project MIGRAWARE, Grant No. 01LG2082C, funded by the German Federal Ministry of Education and Research (BMBF) as part of the programme WASCAL WRAP 2.0, and the research and teaching initiative “A New Focus in English Linguistics: Geolingual Studies”, Grant No. 98 662, funded by Volkswagen Foundation as part of the programme “Momentum—Funding for Recently Tenured Professors”. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions. The implementation of the migration-detection algorithm was published as an openly available R-package MigrationDetectR (https://github.com/dlr-eoc/ukis-migrationdetectr).

ORCID

Johannes Mast http://orcid.org/0000-0001-6595-5834
Marta Sapena https://orcid.org/0000-0003-3283-319X
Martin Mühlbauer https://orcid.org/0000-0003-3849-1143
Carolin Biewer https://orcid.org/0000-0002-3797-1586
Hannes Taubenböck https://orcid.org/0000-0003-4360-9126

REFERENCES

Afolabi, O. J., Oluwaji, O. A., & Fashola, O. K. (2017). Socio-economic impact of road traffic congestion on urban mobility: A case study of Ikeja Local Government Area of Lagos State, Nigeria. *Pacific Journal of Science and Technology*, *18*(2), 246–255.

Akanle, O., Fayehun, O., & Oyelakin, S. (2021). The information communication technology, social media, international migration and migrants' relations with Kin in Nigeria. *Journal of Asian and African Studies*, *56*(6), 1212–1225. https://doi.org/10.1177/0021909620960148

Armstrong, C., Poorthuis, A., Zook, M., Ruths, D., & Soehl, T. (2021). Challenges when identifying migration from geo-located Twitter data. *EPJ Data Science*, *10*(1), 1. https://doi.org/10.1140/epjds/s13688-020-00254-7

Bastian, M., Hayes, M., Vaughan, W., Shah, S., Skomoroch, P., Kim, H., Uryasev, S., & Lloyd, C. (2014). LinkedIn skills: Large-scale topic extraction and inference. *In Proceedings of the 8th ACM Conference on Recommender Systems* (pp. 1–8). Association for Computing Machinery (ACM). https://doi.org/10.1145/2645710.2645729

Bloch, R., Fox, S., Monroy, J., & Ojo, A. (2015). *Urbanisation and urban expansion in Nigeria* (Vol. *73*). ICF International.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. *Information Technology for Development*, *18*(2), 107–125. https://doi.org/10.1080/02681102.2011.643209

Bosco, C., Grubanov-Boskovic, S., Iacus, S. M., Minora, U., Sermi, F., & Spyratos, S., (2022). *Data innovation in demography, migration and human mobility* (EUR 30907 EN). Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/027157

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Calderón, C. A., de la Vega, G., & Herrero, D. B. (2020). Topic modeling and characterization of hate speech against immigrants on Twitter around the emergence of a far-right party in Spain. *Social Sciences*, *9*(11), 188. https://doi.org/10.3390/socsci9110188

Central Intelligence Agency. (2022, June 14). *Nigeria—The world factbook*. https://www.cia.gov/the-world-factbook/countries/nigeria/

Charmaz, K. (2006). *Constructing grounded theory*. Sage Publications.

Chi, G., Lin, F., Chi, G., & Blumenstock, J. (2020). A general approach to detecting migration events in digital trace data. *PLoS ONE*, *15*(10), e0239408. https://doi.org/10.1371/journal.pone.0239408

Cleveland, W. S., & Loader, C. (1996). Smoothing by local regression: Principles and methods. In: W. Härdle & M. G. Schimek (Eds.), *Statistical theory and computational aspects of smoothing* (pp. 10–49). Physica HD. https://doi.org/10.1007/978-3-642-48425-4_2

Dekker, R., & Engbersen, G. (2014). How social media transform migrant networks and facilitate migration. *Global Networks*, *14*(4), 401–418. https://doi.org/10.1111/glob.12040

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

Didia, D., & Tahir, S. (2022). Enhancing economic growth and government revenue generation in Nigeria: The role of diaspora remittances. *The Review of Black Political Economy*, *49*(2), 175–202. https://doi.org/10.1177/00346446211025647

Economic Intelligence Unit, M. of E. P., & B. (2013). *The socio-economic costs of traffic congestion in Lagos* [Working Paper Series, 2].

Efthimion, P. G., Payne, S., & Proferes, N. (2018). Supervised machine learning bot detection techniques to identify social Twitter bots. *SMU Data Science Review*, *1*(2), 71.

Egghe, L., & Leydesdorff, L. (2009). The relation between Pearson's correlation coefficient r and Salton's cosine measure. *Journal of the American Society for Information Science and Technology*, *60*(5), 1027–1036.

Fiallos, A., & Jimenes, K. (2019). Using Reddit data for multi-label text classification of Twitter users interests. *In 2019 Sixth International Conference on EDemocracy & EGovernment (ICEDEG)* (pp. 324–327). Institute for Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/ICEDEG.2019.8734365

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017). Using Twitter data to estimate the relationship between short-term mobility and long-term migration. *In Proceedings of the 2017 ACM on Web Science Conference* (pp. 103–110). Association for Computing Machinery (ACM). https://doi.org/10.1145/3091478.3091496

Forenbacher, I., Husnjak, S., Cvitić, I., & Jovović, I. (2019). Determinants of mobile phone ownership in Nigeria. *Telecommunications Policy*, *43*(7), 101812. https://doi.org/10.1016/j.telpol.2019.03.001

Fussell, E., Hunter, L. M., & Gray, C. L. (2014). Measuring the environmental dimensions of human migration: The demographer's toolkit. *Global Environmental Change*, *28*, 182–191. https://doi.org/10.1016/j.gloenvcha.2014.07.001

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(59), 1–35.

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, *49*(2), 1–41. https://doi.org/10.1145/2938640

Gollin, D., Blanchard, P., & Kirchberger, M. (2021). *Perpetual motion: Human mobility and spatial frictions in three African countries* (SSRN Scholarly Paper No. 3960245). https://papers.ssrn.com/abstract=3960245

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Heidenreich, T., Eberl, J.-M., Lind, F., & Boomgaarden, H. (2020). Political migration discourses on social media: A comparative perspective on visibility and sentiment across political Facebook accounts in Europe. *Journal of Ethnic and Migration Studies*, *46*(7), 1261–1280. https://doi.org/10.1080/1369183X.2019.1665990

Hoffmann, R., Šedová, B., & Vinke, K. (2021). Improving the evidence base: A methodological review of the quantitative climate migration literature. *Global Environmental Change*, *71*, 102367. https://doi.org/10.1016/j.gloenvcha.2021.102367

Huang, B., & Carley, K. M. (2019). A large-scale empirical study of geotagging behavior on Twitter. *In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 365–373). Association for Computing Machinery (ACM). https://doi.org/10.1145/3341161.3342870

Hübl, F., Cvetojevic, S., Hochmair, H., & Paulus, G. (2017). Analyzing refugee migration patterns using geo-tagged tweets. *ISPRS International Journal of Geo-Information*, *6*(10), 302. https://doi.org/10.3390/ijgi6100302

Hughes, C., Zagheni, E., Abel, G. J., Sorichetta, A., Wi'sniowski, A., Weber, I., & Tatem, A. J. (2016). *Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data: Feasibility study on inferring (labour) mobility and migration in the European union from big data and social media data*. European Commission. https://doi.org/10.2767/61617

Igwe, P. A., Obatolu, A. D. A., Nwajiuba, C. A., Egbo, O. P., Ogunnaike, O. O., & Nwekpa, K. C. (2021). The glocalisation of sports: A study of the influence of European Football Leagues on Nigerian society. *European Journal of International Management*, *15*(2–3), 247–265. https://doi.org/10.1504/EJIM.2021.113244

International Organization for Migration (IOM) & Displacement Tracking Matrix (DTM). (2022). *Nigeria flood map overview (as of October 2022)*. https://displacement.iom.int/

Johnson, I. L., Sengupta, S., Schöning, J., & Hecht, B. (2016). The geography and importance of localness in geotagged social media. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 515–526). Association for Computing Machinery (ACM). https://doi.org/10.1145/2858036.2858122

Kayastha, T., Gupta, P., & Bhattacharyya, P. (2021). BERT based adverse drug effect tweet classification. In M. Arjun, K. Ari, M.-E. Antonio, A. A. Mohammed, A. Ilseyar, M. Zulfat, F.-M. Eulalia, L. L. Salvador, F. Ivan, O. Karen, W. Davy, T. Elena, S. Abeed, M. B. Juan, K. Martin, & G.-H. Graciela (Eds.), *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task* (pp. 88–90). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.smm4h-1.15

Kemp, S. (2015). *Digital 2015 Nigeria*. Datareportal. https://datareportal.com/reports/digital-2015-nigeria

Kemp, S. (2019). *Digital 2019 Nigeria*. Datareportal. https://datareportal.com/reports/digital-2019-nigeria

Khatua, A., & Nejdl, W. (2021). Struggle to settle down! Examining the voices of migrants and refugees on Twitter platform. *In Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 95–98). Association for Computing Machinery (ACM). https://doi.org/10.1145/3462204.3481773

Kim, J., Sîrbu, A., Giannotti, F., & Gabrielli, L. (2020). Digital footprints of international migration on Twitter. In M. R. Berthold, Ad. Feelders, & G. Krempl (Eds.), *Advances in intelligent data analysis XVIII* (pp. 274–286). Springer. https://doi.org/10.1007/978-3-030-44584-3_22

Kim, J., Sîrbu, A., Giannotti, F., & Rossetti, G. (2021). Characterising different communities of Twitter users: Migrants and natives. *In International Conference on Complex Networks and Their Applications*. http://arxiv.org/abs/2103.03710

Kim, J., Sîrbu, A., Giannotti, F., Rossetti, G., & Rapoport, H. (2022). Origin and destination attachment: Study of cultural integration on Twitter. *EPJ Data Science*, *11*(1), 55.

Kirchberger, M. (2021). Measuring internal migration. *Regional Science and Urban Economics*, *91*, 103714. https://doi.org/10.1016/j.regsciurbeco.2021.103714

Kirwin, M., & Anderson, J. (2018). *Identifying the factors driving West African Migration* [West African Papers No. 17; West African Papers, Vol. 17]. OECD. https://doi.org/10.1787/eb3b2806-en

Kochupillai, M., Kahl, M., Schmitt, M., Taubenböck, H., & Zhu, X. X. (2022). Earth observation and artificial intelligence: Understanding emerging ethical issues and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, *10*(4), 90–124.

Kounadi, O., & Resch, B. (2018). A geoprivacy by design guideline for research campaigns that use participatory sensing data. *Journal of Empirical Research on Human Research Ethics*, *13*(3), 203–222. https://doi.org/10.1177/1556264618759877

Kruspe, A., Häberle, M., Hoffmann, E. J., Rode-Hasinger, S., Abdulahhad, K., & Zhu, X. X. (2021). Changes in Twitter geolocations: Insights and suggestions for future usage. In X. Wei, R. Alan, B. Tim, & R. Afshin (Eds.), *Proceedings of the 2021 EMNLP Workshop W-NUT* (pp. 212–221). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2108.12251

Lamanna, F., Lenormand, M., Salas-Olmedo, M. H., Romanillos, G., Gonçalves, B., & Ramasco, J. J. (2018). Immigrant community integration in world cities. *PLoS ONE*, *13*(3), e0191612. https://doi.org/10.1371/journal.pone.0191612

Lee, J.-S., & Nerghes, A. (2018). Refugee or migrant crisis? Labels, perceived agency, and sentiment polarity in online discussions. *Social Media + Society*, *4*(3), 205630511878563. https://doi.org/10.1177/2056305118785638

Leydesdorff, L. (2005). Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information*

Science and Technology, *56*(7), 769–772. https://doi.org/10.1002/asi.20130

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

Malik, M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population bias in geotagged tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(4), 18–27. https://doi.org/10.1609/icwsm.v9i4.14688

Masakhane, ∀ (2022). *Masakhane*. Masakhane. https://www.masakhane.io/

Mau, S., Gülzau, F., Laube, L., & Zaun, N. (2015). The global mobility divide: How visa policies have evolved over time. *Journal of Ethnic and Migration Studies*, *41*(8), 1192–1213. https://doi.org/10.1080/1369183X.2015.1005007

Mazzoli, M., Diechtiareff, B., Tugores, A., Wives, W., Adler, N., Colet, P., & Ramasco, J. J. (2020). Migrant mobility flows characterized with digital data. *PLoS ONE*, *15*(3), e0230264. https://doi.org/10.1371/journal.pone.0230264

McAuliffe, M., Kitimbo, A., Goossens, A. M., & Ullah, A. A. (2018). *World Migration Report 2018—Chapter 7—Understanding migration journeys from migrants' perspectives* (World Migration Report, 2018(1)).

McAuliffe, M., & Ruhs, M. (2018). *World Migration Report 2018—Introduction* (World Migration Report, 1).

McCreadie, R., Buntain, C. L., & Soboroff, I. (2019). TREC incident streams: Finding actionable information on social media. *International Conference on Information Systems for Crisis Response and Management (ISCRAM 2019), Valencia, Spain, 19-22 May 2019*. pp. 691–705.

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English Tweets. In L. Qun, & S. David (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.2

NiDCOM. (2021). About NiDCOM – Nigerians in diaspora commission. About NiDCOM. https://nidcom.gov.ng/about-nidcom/

NOI Polls Ltd. (2015). *US and UK top list of countries most Nigerians abroad reside in*. https://noi-polls.com/us-and-uk-top-list-of-countries-most-nigerians-abroad-reside-in-key-reason-for-migration-is-for-economic-opportunities/

Ojedokun, U. A., Ogunleye, Y. O., & Aderinto, A. A. (2021). Mass mobilization for police accountability: The case of Nigeria's #EndSARS protest. *Policing: A Journal of Policy and Practice*, *15*(3), 1894–1903. https://doi.org/10.1093/police/paab001

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13. https://doi.org/10.3389/fdata.2019.00013

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing & Management*, *57*(4), 102250. https://doi.org/10.1016/j.ipm.2020.102250

Orekan, G. (2010). Language policy and educational development in Africa: The case of Nigeria. *Scottish Languages Review*, *21*, 17–26.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., & Barabási, A.-L. (2015). Returners and explorers dichotomy in human mobility. *Nature Communications*, *6*(1), 8166. https://doi.org/10.1038/ncomms9166

Petutschnig, A., Havas, C. R., Resch, B., Krieger, V., & Ferner, C. (2020). Exploratory spatiotemporal language analysis of geo-social network data for identifying movements of refugees. *GI_Forum*, *1*, 137–152. https://doi.org/10.1553/giscience2020_01_s137

Princewill, N., & Busari, S., CNN. (2021, June). *Nigeria bans Twitter after company deletes president Buhari's tweet*. CNN. https://www.cnn.com/2021/06/04/africa/nigeria-suspends-Twitter-operations-intl/index.html

Quercia, D., Askham, H., & Crowcroft, J. (2012). TweetLDA: Supervised topic classification and link prediction in Twitter. *In Proceedings of the 3rd Annual ACM Web Science Conference on—WebSci '12* (pp. 247–250). Association for Computing Machinery (ACM). https://doi.org/10.1145/2380718.2380750

Rampazzo, F., Bijak, J., Vitali, A., Weber, I., & Zagheni, E. (2021). A framework for estimating migrant stocks using digital traces and survey data: An application in the United Kingdom. *Demography*, *58*(6), 2193–2218. https://doi.org/10.1215/00703370-9578562

Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In S. Donia, B. Nuria, & Z. Chengqing (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6838–6855). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.603

Reips, U.-D., & Buffardi, L. E. (2012). Studying migrants with the help of the Internet: Methods from psychology. *Journal of Ethnic and Migration Studies*, *38*(9), 1405–1424. https://doi.org/10.1080/1369183X.2012.698208

Rowe, F., Mahony, M., Graells-Garrido, E., Rango, M., & Sievers, N. (2021). Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. *Data & Policy*, *3*, e36. https://doi.org/10.1017/dap.2021.38

Schewel, K. (2020). Understanding immobility: Moving beyond the mobility bias in migration studies. *International Migration Review*, *54*(2), 328–355. https://doi.org/10.1177/0197918319831952

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., Guidotti, R., Bertoli, S., Kim, J., Muntean, C. I., Pappalardo, L., Passarella, A., Pedreschi, D., Pollacci, L., Pratesi, F., & Sharma, R. (2021). Human migration: The big data perspective. *International Journal of Data Science and Analytics*, *11*(4), 341–360. https://doi.org/10.1007/s41060-020-00213-5

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2018). *Migration data using social media: A European perspective*. Publications Office of the European Union. https://hdl.handle.net/21.11116/0000-0004-7C55-2

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., & Menczer, F. (2016). The DARPA Twitter bot challenge. *Computer*, *49*(6), 38–46. https://doi.org/10.1109/MC.2016.183

Taubenböck, H., Staab, J., Zhu, X. X., Geiß, C., Dech, S., & Wurm, M. (2018). Are the poor digitally left behind? Analyzing urban divides using remote sensing and Twitter data. *ISPRS International Journal of Geo-Information*, *7*(8), 304.

Torch Contributors. (2019). *CrossEntropyLoss—PyTorch 1.11.0 documentation*. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

Townsend, L., & Wallace, C. (2016). *Social media research: A guide to ethics* (pp. 1–16). University of Aberdeen.

Twitter Inc. (2022a). *Twitter API for academic research | products*. https://developer.Twitter.com/en/products/Twitter-api/academic-research

Twitter Inc. (2022b, June). *Batch compliance*. https://developer.Twitter.com/en/docs/Twitter-api/compliance/batch-compliance/introduction

United Nations Department of Economic and Social Affairs. (1998). *Recommendations on statistics of international migration. Revision 1*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wang, Z., Hale, S. A., Adelani, D., Grabowicz, P. A., Hartmann, T., Flöck, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. *In The World Wide Web Conference* (pp. 2056–2067). Association for Computing Machinery (ACM). https://doi.org/10.1145/3308558.3313684

Willekens, F. (2019). Evidence-based monitoring of international migration flows in Europe. *Journal of Official Statistics*, *35*(1), 231–277. https://doi.org/10.2478/jos-2019-0011

Willekens, F., Massey, D., Raymer, J., & Beauchemin, C. (2016). International migration under the microscope. *Science*, *352*(6288), 897–899. https://doi.org/10.1126/science.aaf6545

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/003803851770814

World Bank. (2022). *Remittances brave global headwinds. Special focus: Climate migration*. Knomad. https://www.knomad.org/sites/default/files/publication-doc/migration_and_development_brief_37_nov_2022.pdf

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from Twitter data. *In Proceedings of the 23rd International Conference on World Wide Web* (pp. 439–444). Association for Computing Machinery (ACM). https://doi.org/10.1145/2567948.2576930

Zhu, X. X., Wang, Y., Kochupillai, M., Werner, M., Haberle, M., Hoffmann, E. J., Taubenbock, H., Tuia, D., Levering, A., Jacobs, N., Kruspe, A., & Abdulahhad, K. (2022). Geoinformation harvesting from social media data: A community remote sensing approach. *IEEE Geoscience and Remote Sensing Magazine*, *10*(4), 150–180. https://doi.org/10.1109/MGRS.2022.3219584

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mast, J., Sapena, M., Mühlbauer, M., Biewer, C., & Taubenböck, H. (2023). The migrant perspective: Measuring migrants' movements and interests using geolocated tweets. *Population, Space and Place*, e2732. https://doi.org/10.1002/psp.2732